J. SYSTEMS SOFTWARE 1992; 17:7–14

Human Agency and Responsible Computing: Implications for Computer System Design

Batya Friedman
Mills College, Oakland, California

Peter H. Kahn, Jr.
University of Houston, Houston, Texas

To understand and promote responsible computing, this paper highlights the importance of analyses based on human agency. We first examine whether computers can be moral agents. Then we draw on research in human factors, cognitive science, and instructional technology to examine how three types of computing practices can be problematic from the perspective of human agency. The first involves anthropomorphizing a computational system, the second, delegating decision making to a computational system, and the third, delegating instruction to a computational system. Throughout this discussion, we provide alternative design goals and methods by which responsible computing can be enhanced as a shared vision and practice within the computing community.

Societal interest in responsible computing perhaps most often arises in response to harmful consequences that can result from computing. For instance, consider the frustration and economic loss incurred by individuals and businesses whose computer systems have been infected by the Internet worm or other computer viruses. Or consider the physical suffering and death of the cancer patients who were overradiated by Therac-25, or of civilians accidentally bombed in the Persian Gulf war by "smart" missiles gone astray. Largely in reaction to events like these, we have in recent years seen a surge of interest in preventing or at least minimizing such harmful consequences. But if responsible computing is to be understood as something more than a form of damage control, how are we to understand the term? Moreover, how can responsible computing be promoted within the computing community?

In this article, we address these questions by highlighting the importance of analyses based not only on consequences of acts, but agency—on what and why some things can be held morally responsible for action. We shall first examine whether computers can be such things. While our discussion here will be largely philosophical (and somewhat condensed as each piece may well be familiar to the reader), a compelling position on whether computers can be moral agents provides an important starting point for our central task. We seek to understand how, from the moral perspective, we should conceive of the human relationship to computational systems, and to provide sketches of how to build on that conception to promote responsible computing through system design. To this end, we will examine how three types of computing practices can be problematic from the perspective of human agency. The first involves anthropomorphizing a computational system, the second, delegating decision making to a computational system, and the third, delegating instruction to a computational system.

CAN COMPUTERS BE MORAL AGENTS?

To understand the place and urgency of the question of whether computers can be moral agents, consider the issues raised by computer-based closed-loop drug administration systems. In critical care medicine, these automated systems are designed to monitor and, when necessary, adjust the administration of a variety of drugs for patients in an intensive care unit. Such

Address correspondence to Batya Friedman, Clare Boothe Luce Asst. Professor of Computer Science, Dept. of Mathematics and Computer Science, Colby College, Waterville, ME 04901.

655 Avenue of the Americas, New York, NY 10010

0164-1212/92/$5.00

computer-based systems are touted for their increased effectiveness over human-administered drug therapy, for their safety, and for their usefulness in reducing nursing demands [1, 2]. However, these systems, although currently recommended for use, pose ethical problems. For instance, Snapper [3] suggests that such a computer-based system "may not be able to check as many variables as could a doctor at the bedside and so may administer the wrong drug when a doctor would administer the correct drug" (p. 289). Or such a computer-based system may not be programmed to account for a particular atypical case, and so may administer the wrong drug when an experienced doctor would administer the correct drug. In such situations, can the computer-based closed-loop drug administration system itself be held, even in part, morally responsible for the decision to administer a wrong drug? Stated more generally, can a computational system be a moral agent and thus be held morally responsible for a decision?

Toward addressing this question, consider two cases. While hiking in the mountains, Y is crushed by a falling boulder and killed. In case one, the boulder was dislodged by a slight shifting and settling of the land on which it balanced. In case two, the same boulder was dislodged by a push from X, who desired to kill Y, believed the push would cause the boulder to fall on Y, understood that a boulder of such weight would kill Y, and freely chose to perform the act. The cases are the same in that some "thing" caused a boulder to fall, killing Y. But the cases are fundamentally different in that only in the second case was the act of dislodging the boulder the result of an intentional act. This distinction between the cases highlights a philosophical position that for a thing to be held morally responsible it must be capable of intentionality, which, at a minimum, refers to the capability of having or experiencing beliefs, desires, understandings, intentions, and volition [4, 5]. Given, then, that people but not land are usually considered psychologically capable of intentional states, only X (but not the land) could be held morally responsible for the death of Y. Moving now to the question at hand: If we accept that intentional states are a prerequisite for a thing to be held morally responsible, then a subset of the above question—can a computational system be a moral agent and thus be held morally responsible for a decision?—can be framed as follows: Can a computational system be considered to have intentional states?

Much of the literature in artificial intelligence would have us think it so, or at least have us think it possible. One classic framing of this position can be traced to McCarthy [6], who proposed some decades ago that machine intentionality is equivalent to human intentionality. For example, in one of his well-known analogies, McCarthy claimed that a thermostat has beliefs (of whether a room is too hot or cold) that lead to intended action (turning the heater off or on). Such intentional states, according to McCarthy, are equivalent to that of a human who can have beliefs (of whether a room is too hot or cold) that lead to intended action (taking a sweater off or putting one on).

There are two ways to understand such a position that equate machine intentionality with human intentionality. Both will frame our analyses of computer system design. In the first, whatever we may call intentionality and think we may experience in terms of feelings, beliefs, understandings, free will, or an underlying sense of self or personhood are epiphenomenal, meaning such experiences play no authentic causal role in our actions. In the second, it is claimed that machines have (or with increased technological advancements will have) psychological states similar or identical to those which comprise human intentionality (in a nonepiphenomenal sense). In other words, the first reduces humans to the status of computers, while the second raises computers to the status of humans.

Both ways of understanding are problematic. Granted, it may be that humans ultimately will never be able to prove conclusively that what we take to be intentionality is not epiphenomenal, for the position draws on a radical skepticism that calls into question every means we might have to undermine it. The argument is similar to one that charges that you, the reader, are nothing but a brain in a vat [7], prodded at this very moment with electrical stimulation to induce you to think that you are reading this essay, and that you have the thoughts, feelings, and experiences that you do. Anything you might try to say to counterthis position (e.g., "I think therefore I am," or less formally, "But I know deep down inside myself that that is not true") can be counterargued with the claim that your knowledge has simply been induced by electrical stimulation.

It is a far cry, however, from not being able to prove this position conclusively false to believing, with good reason, that it is true. Phenomenologically, humans experience intentional states and believe they have beliefs, understandings, and free will. If such intentionality is epiphenomenal, it is difficult to understand biologically and psychologically why or how it ever originated within our species. Moreover, it would be virtually impossible to live in this world without taking our intentional states seriously. We would, for instance, have to abandon such beliefs that a difference exists between accidental and intended harm (since the belief in intended action is epiphenomenal), that persons can lose weight, climb hills, or read books if they so choose (since the belief in free choice is epiphenomenal), and so on for the countless intentional states that pervade our lives. Indeed, even the desire to under-

J. SYSTEMS SOFTWARE
1992; 17:7-14

stand how intentionality is epiphenomenal presupposes a validity to intentionality, to such psychological constructs as desire and understanding that lead to the intended action to provide an alternative explanation. The point here is that without positive evidence to the contrary, which this first position based on a radical skepticism does not provide, humans have good reason to believe that human intentionality plays an authentic causal role in our actions.

The second way of understanding the proposition that equates human intentionality with machine intentionality, that machines have states similar or identical to those which comprise human intentionality, has been substantively critiqued by Searle [8]. His Chinese room argument is well known:

> Consider a language you don't understand. In my case, I do not understand Chinese. To me Chinese writing looks like so many meaningless squiggles. Now suppose I am placed in a room containing baskets full of Chinese symbols. Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols. The rules identify the symbols entirely by their shapes and do not require that I understand any of them. Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols. Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base" . . . . Now suppose that the rule book is written in such a way that my "answers" to the "questions" are indistinguishable from those of a native Chinese speaker . . . . All the same, I am totally ignorant of Chinese. And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols. Like a computer, I manipulate symbols, but I attach no meaning to the symbols. (p. 26)

In other words, because computational systems are purely formal (syntax), and because purely formal systems have no means to generate intentionality (semantics), computational systems do not have intentionality.

Searle's position has generated a great deal of debate, including 26 commentaries, since it appeared in 1980 [9], and continued more recently by Churchland and Churchland [10]. While this is not the place to review the many arguments and counterarguments in the debate, in our view and the view of others, Searle has defended his position well against his critics. This is not to say that minds and their intentional states might not someday be realized in materials or structures other than biological brains. But it is to say that computers as we can conceive of them today are not such materials or structures.

Thus we have argued, however briefly, for three propositions: 1) intentionality is a necessary condition of moral agency; 2) we can, with confidence, believe that human intentionality plays an authentic causal role in our actions; and 3) a computer system as we can conceive of it today in material and structure cannot have intentionality. From these three propositions, it follows that humans, but not computational systems, are capable of being moral agents, and that humans, but not computational systems, are capable of being morally responsible for computer-mediated actions and consequences.

DESIGN TO SUPPORT HUMAN AGENCY AND RESPONSIBLE COMPUTING

Based on this line of reasoning, we propose that responsible computing often depends on humans' clear understanding that humans are capable of being moral agents and that computational systems are not. However, as anticipated by the above discussion, this understanding can be distorted in one of two ways. In the first type of distortion, the computational system diminishes or undermines the human user's sense of his or her own moral agency. In such systems, human users are placed into largely mechanical roles, either mentally or physically, and frequently have little understanding of the larger purpose or meaning of their individual actions. To the extent that humans experience a diminished sense of agency, human dignity is eroded and individuals may consider themselves to be largely unaccountable for the consequences of their computer use. Conversely, in the second type of distortion the computational system masquerades as an agent by projecting intentions, desires, and volition. To the extent that humans inappropriately attribute agency to such systems, humans may well consider the computational systems, at least in part, to be morally responsible for the effects of computer-mediated or computer-controlled actions.

Accordingly, to support humans' responsible use of computational systems, system design should strive to minimize both types of distortion. That is, system design should seek to protect the moral agency of humans and to discourage in humans a perception of moral agency in the computational system. How might design practices achieve these goals? Given that little research exists that addresses this question directly, we seek to provide some initial sketches by examining three types of computer practices.

Anthropomorphizing the Computational System

Anthropomorphic metaphors can be found in some of the definitions and goals for interface design. For

example, some interfaces are designed to "use the process of human-human communication as a model for human-computer interaction" ([11], p. 86), to "interact with the user similar to the way one human would interact with another" ([11], p. 87), or to be "intelligent" where intelligence is based on a model of human intelligence. When such anthropomorphic metaphors become embedded in the design of a system, the system can fall prey to the second type of distortion by projecting human agency onto the computational system.

Moreover, even in unsophisticated designs of this type, there is some evidence that people do attribute agency to the computational system. For example, Weizenbaum [12] reported that some adults interacted with his computer program DOCTOR with great emotional depth and intimacy, "conversing with the computer as if it were a person" (p. 7). In a similar vein, some of the children Turkle [13] interviewed about their experiences with an interactive computer game called Merlin that played Tic-Tac-Toe attributed psychological (mental) characteristics to Merlin. For example, children sometimes accused Merlin of cheating, an accusation that includes a belief that the computer has both the intention and desire to deceive. In another example, Rumelhart and Norman [14] attempted to teach novices to use an editing program by telling the novices that the system was like a secretary. The novices drew on this human analogy to attribute aspects of a secretary's intelligence to the editing system and assumed (incorrectly) that the system would be able to understand whether they intended a particular string of characters to count as text or as commands.

While these examples of human attribution of agency to computational systems have largely benign consequences, this may not always be the case. Consider Jenkins' [15] human factors experiment that simulated a nuclear power plant failure. In the experiment, nuclear power plant operators had access to an expert system to aid them in responding to the plant failure. Although previously instructed on the expert system's limitations, the "operators expected that the expert system implemented in the computer 'knew' about the failures of the cooling system without being told. The system [however] was neither designed nor functioned as an automatic fault recognition system" (p. 258). Jenkins attributed this overestimation of the system's capabilities to the power plant operators' expectations for the expert system to know certain information, presumably the type of information that any responsible human expert would know or attempt to find out in that situation.

Because nonanthropomorphic design does not encourage people to attribute agency to the computational system, such designs can better support responsible computing. To clarify what such design looks like in practice, consider the possibilities for interface design. Without ever impersonating human agency, interface design can appropriately pursue such goals as learnability, ease and pleasure of use, clarity, and quick recovery from errors. In addition, nonanthropomorphic interface design can employ such techniques as novel pointing devices, nonanthropomorphic analogies, speech input and output, and menu selection. Or consider the characteristics of another plausible technique: direct manipulation. According to Jacob [16], direct manipulation refers to a user interface in which the user "seems to operate directly *on* the objects in the computer rather than carrying on a dialogue *about* them" (p. 166). For example, the Xerox Star desktop manager adapted for systems such as the Apple Macintosh uses images of standard office objects (e.g., files, folders, and trash cans) and tasks to represent corresponding objects and functions in the editing system [17]. In this environment, disposing of a computer file is achieved by moving the image of the file onto the image of the trash can, akin to disposing of a paper file by physically placing the file in a trash can. There is no ambiguity in this direct manipulation interface as to who is doing the acting (the human user) and what the user is acting upon (objects in the computational system). The defining characteristics of direct manipulation suggest that this technique would not lead to projecting human agency onto the system. This is because direct manipulation involves physical action on an object as opposed to social interaction with an other as an underlying metaphor. Additionally, direct manipulation seeks to have the human user directly manipulate computational objects, thereby virtually eliminating the possibility for the human user to perceive the computer interface as an intermediary agent.

Nonanthropomorphic design considerations fit within a larger vision for interface design that is already part of the field. For example, Shneiderman [18] draws on Weizenbaum [12] to advocate design that "sharpen[s] the boundaries between people and computers . . . [for] human-human communication is a poor model for human-computer interaction" (p. 434). More recently, Shneiderman [19] writes that "when an interactive system is well designed, it almost disappears, enabling the users to concentrate on their work or pleasure" (p. 169). Winograd and Flores [20] similarly advocate the design of nonanthropomorphic computer tools that provide a transparent interaction between the user and the resulting action. "The transparency of interaction is of utmost importance in the design of tools, including computer systems, but it is not best achieved by attempting to mimic human faculties" (p. 194). When a transparent interaction is achieved, the user is freed from the details of using the tool to focus on the task at

hand. The shared vision here is for the interface to "disappear," not to intercede in the guise of another "agent" between human users and the computational system.

Delegating Decision Making to Computational Systems

When delegating decision making to computational systems, both types of distortions can occur. The discussion that follows examines these distortions in the context of the APACHE system [21, 22]. More generally, however, similar analyses could be applied to other computer-based models and knowledge-based systems such as MYCIN [23] or the Authorizer's Assistant used by the American Express Corporation [24].

APACHE is a computer-based model implemented but not yet used clinically that determines when to withdraw life support systems from patients in intensive care units. Consider the nature of the human–computer relationship if APACHE, used as a closed-loop system, determines that life support systems should be withdrawn from a patient, and then turns off the life support systems. In ending the patient's life the APACHE system projects a view of itself to the medical personnel and the patient's family as a purposeful decision maker (the second type of distortion). Simultaneously, the system allows the attending physician and critical care staff to distance or numb themselves from the decision making process about when to end another human's life (the first type of distortion).

Now, in actuality, at least some of the researchers developing APACHE do not recommend its use as a closed-loop system, but as a consultation system, one that recommends a course of action to a human user who may or may not choose to follow the recommendation [21]. These researchers write: "Computer predictions should never dictate clinical decisions, as very often there are many factors other than physiologic data to be considered when a decision to withdraw therapy is made" (p. 1096). Thus, used as a consultation system, APACHE functions as a tool to aid the critical care staff with making difficult decisions about the withdrawal of therapy. Framed in this manner, the consultation system approach seems to avoid the distortions of human agency described above: the consultation system does not mimic purposeful action or inappropriately distance the medical staff from making decisions about human life and death.

In practice, however, the situation can be more complicated. Most human activity, including the decision by medical personnel to withdraw life support systems, occurs in a web of human relationships. In some circumstances, because a computational system is embedded in a complex social structure human users may experience a diminished sense of moral agency. Let us imagine, for instance, that APACHE is used as a consultation system. With increasing use and continued good performance by APACHE, it is likely that the medical personnel using APACHE will develop increased trust in APACHE's recommendations. Over time, these recommendations will carry increasingly greater authority within the medical community. Within this social context, it may become the practice for critical care staff to act on APACHE's recommendations somewhat automatically, and increasingly difficult for even an experienced physician to challenge the "authority" of APACHE's recommendation, since to challenge APACHE would be to challenge the medical community. But at this point the open-loop consultation system through the social context has become, in effect, a closed-loop system wherein computer prediction dictates clinical decisions.

Such potential effects point to the need to design computational systems with an eye toward the larger social context, including long-term effects that may not become apparent until the technology is well situated in the social environment. Participatory design methods offer one such means [25, 26]. Future users, who are experienced in their respective fields, are substantively involved in the design process. As noted at a recent conference [27], Thoresen worked with hospital nurses to design a computer-based record-keeping system. In the design process, nurses helped to define on a macro level what institutional problems the technology would seek to solve, and on a micro level how such technological solutions would be implemented. From the perspective of human agency, such participatory design lays the groundwork for users to see themselves as responsible for shaping the system's design and use.

Delegating Instruction to Computational Systems

Instructional technology programs that deliver systematically designed computer-based courseware to students can suffer from the first type of distortion—computer use that erodes the human user's sense of his or her own agency. Often absent from this type of instructional technology is a meaningful notion of the student's responsibility for learning. Johnsen and Taylor [28] have discussed this problem in a paper aptly titled "At cross-purpose: instructional technology and the erosion of personal responsibility." According to Johnsen and Taylor, instructional technology "define[s] responsibility operationally in the context of means/ends rationality. The singular responsibility for a student's education becomes identified with the success of the program" (p. 9). They further point to the logical conclusion of this educational view for students, parents, teachers, and government: failure to educate comes to mean that

the instructional technology failed to teach, not that students failed to learn.

As an example of this type of instructional technology, consider how the GREATERP intelligent tutoring system (described in [29]) for novice programmers in LISP handles students' errors. When GREATERP determines the student has entered "incorrect" information, the tutor interrupts the student's progress toward the student's proposed solution (viable or not) and forces the student to backtrack to the intelligent tutor's "correct" solution. Thus GREATERP assumes responsibility not only for student learning but also for preventing student errors along the way and for the process of achieving a solution. In so doing, this intelligent tutoring system—and other comparable instructional technology programs—can undermine the student's sense of his or her own agency and responsibility for the educational endeavor.

In contrast, other educational uses of computing promote students' sense of agency and active decision making. For example, just as consultation systems can to some degree place responsibility for decision making on the human user, so educational uses of computer applications software (e.g., word processors, spreadsheets, data bases, microcomputer-based labs) can place responsibility for learning on the student. With computer applications students determine when the applications would be useful and for what purposes, and evaluate the results of their use. Moreover, the social organization of school computer use can contribute to students' understanding of responsible computing. As with participatory design, consider the value of student participation in creating the policies that govern their own school computer use. For example, as discussed in a recent article by Friedman [30], students can determine the privacy policy for their own electronic mail at school. To establish such a privacy policy, "students must draw on their fundamental understandings of privacy rights to develop specific policies for this new situation. In turn, circumstances like these provide opportunities for students not only to develop morally but to make decisions about a socially and computationally powerful technology, and thus to mitigate a belief held by many people that one is controlled by rather than in control of technology." Through such experiences, students can learn that humans determine how computer technology is used and that humans bear responsibility for the results of that use.

CONCLUSION

We argued initially that humans, but not computers (as they can be conceived today in material and structure), are or could be moral agents. Based on this view, we identified two broad approaches by which computer system design can promote responsible computer use. Each approach seeks to minimize a potential distortion between human agency and computer activity. First, computational systems should be designed in ways that do not denigrate the human user to machine-like status. Second, computational systems should be designed in ways that do not impersonate human agency by attempting to mimic intentional states. Both approaches seek to promote the human user's autonomous decision making in ways that are responsive to and informed by community and culture.

What we have provided, of course, are only broad approaches and design sketches. But if we are correct that human agency is central to most endeavors that seek to understand and promote responsible computing, then increased attention should be given to how the human user perceives specific types of human-computer interactions, and how human agency is constrained, promoted, or otherwise affected by the larger social environment. In such investigations, it is likely that research methods can draw substantively on existing methods employed in the social-cognitive and moral-developmental psychological fields. Methods might include 1) semistructured hypothetical interviews with participants about centrally relevant problems [31-35]; 2) naturalistic and structured observations [36-38]; and 3) semistructured interviews based on observations of the participant's practice [39-41]. Of note, some anthropologists [42] and psychologists [43] working in the area of human factors have with some success incorporated aspects of these methods into their design practices.

A final word needs to be said about the role of moral psychology in the field of computer system design. As increasingly sophisticated computational systems have become embedded in social lives and societal practices, increasing pressure has been placed on the computing field to go beyond purely technical considerations and to promote responsible computing. In response, there has been, understandably, a desire to know the "right" answer to ethical problems that arise, where "right" is understood to mean something like "philosophically justified or grounded." We agree that there is an important place for philosophical analyses in the field. But philosophy seldom tells us how or why problems relevant to a philosophical position involving computing occur in practice, let alone what can most effectively resolve them. Such issues require empirical data that deal substantively with the psychological reality of humans. Thus, by linking our technical pursuits with both philosophical inquiry and moral-psychological research, responsible computing can be enhanced as a

shared vision and practice within the computing community.

REFERENCES

1. P. Bednarski, F. Siclari, A. Voigt, S. Demertzis, and G. Lau, Use of a computerized closed-loop sodium nitroprusside titration system for antihypertensive treatment after open heart surgery, *Crit. Care Med.* 18, 1061–1065 1990).
2. S. McKinley, J. F. Cade, R. Siganporia, O. M. Evans, D. G. Mason, and J. S. Packer, Clinical evaluation of closed-loop control of blood pressure in seriously ill patients, *Crit. Care Med.* 19, 166–170 (1991).
3. J. W. Snapper, Responsibility for computer-based errors, *Metaphilosophy* 16, 289–295 (1985).
4. G. E. Scott, *Moral Personhood: An Essay in the Philosophy of Moral Psychology*, State University of New York Press, Albany, New York, 1990.
5. J. R. Searle, *Intentionality*, Cambridge University Press, Cambridge, England, 1983.
6. J. McCarthy, Ascribing mental qualities to machines, in *Philosophical Perspectives on Artificial Intelligence* (M. Ringle, ed.), Humanities Press, Needham Heights, MA, 1979.
7. A. Zuboff, The story of a brain, in *The Mind's I* (D. R. Hofstadter and D. C. Dennett, eds.), Basic Books, Inc., New York, 1981.
8. J. R. Searle, Is the brain's mind a computer program? *Sci. Am.* 262, 26–31 (1990).
9. J. R. Searle, Minds, brains and programs, *Behav. Brain Sci.* 3, 417–458 (1980).
10. P. M. Churchland and P. S. Churchland, Could a machine think? *Sci. Am.* 262, 32–37 (1990).
11. R. E. Eberts and C. G. Eberts, Four approaches to human computer interaction, in *Intelligent Interfaces: Theory, Research and Design* (P. A. Hancock and M. H. Chignell, eds.), Elsevier Science Publishers, New York, 1989.
12. J. Weizenbaum, *Computer Power and Human Reason*, W. H. Freeman & Company, New York, 1976.
13. S. Turkle, *The Second Self: Computers and the Human Spirit*, Simon & Schuster, New York, 1984.
14. D. E. Rumelhart and D. A. Norman, Analogical processes in learning, in *Cognitive Skills and Their Acquisition* (J. R. Anderson, ed.), Lawrence Erlbaum Associates, Hillsdale, NJ, 1981.
15. J. P. Jenkins, An application of an expert system to problem solving in process control displays, in *Human-Computer Interaction* (G. Salvendy, ed.), Elsevier Science Publishers, New York, 1984.
16. R. J. K. Jacob, Direct manipulation in the intelligent interface, in *Intelligent Interfaces: Theory, Research and Design* (P. A. Hancock and M. H. Chignell, eds.), Elsevier Science Publishers, New York, 1989.
17. D. C. Smith, C. Irby, R. Kimball, W. Verplank, and E. Marslem, Designing the user interface, *Byte* 7, 242–282 (1982).
18. B. Shneiderman, *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1987.
19. B. Shneiderman, Designing the user interface, in *Computers in the Human Context: Information Technology, Productivity, and People* (T. Forester, ed.), The MIT Press, Cambridge, Massachusetts, 1989.
20. T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1986.
21. R. W. S. Chang, B. Lee, S. Jacobs, and B. Lee, Accuracy of decisions to withdraw therapy in critically ill patients: clinical judgment versus a computer model, *Crit. Care Med.* 17, 1091–1097 (1989).
22. J. E. Zimmerman, ed., APACHE III study design: analytic plan for evaluation of severity and outcome, *Crit. Care Med.* 17 (Part 2 Suppl), S169–S221 (1989).
23. E. H. Shortliffe, Medical consultation systems: designing for doctors, in *Designing for Human-Computer Communication* (M. E. Sime and M. J. Coombs, eds.), Academic Press, New York, 1983.
24. C. L. Harris *et al.*, Office automation: making it pay off, in *Computers in the Human Context: Information Technology, Productivity, and People* (T. Forester, ed.), The MIT Press, Cambridge, Massachusetts, 1989.
25. P. Ehn, *Work-oriented Design of Computer Artifacts*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.
26. J. Greenbaum and M. Kyng, eds., *Design at Work: Cooperative Design of Computer Systems*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1990.
27. A. Namioka and D. Schuler, eds., *Proceedings from the Conference on Participatory Design 1990*, Computer Professionals for Social Responsibility, Palo Alto, California, 1990.
28. J. B. Johnsen and W. D. Taylor, At cross-purpose: instructional technology and the erosion of personal responsibility, paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1988.
29. R. Kass, Student modeling in intelligent tutoring systems—implications for user modeling, in *User Models in Dialog Systems* (A. Kobsa and W. Wahlster, eds.), Springer-Verlag, New York, 1989.
30. B. Friedman, Social and moral development through computer use: a constructivist approach, *J. Res. Comput. Educ.* 23: 560–567 (1991).
31. W. Damon, *The Social World of the Child*, Jossey-Bass, San Francisco, 1977.
32. L. Kohlberg, Stage and sequence: the cognitive-developmental approach to socialization, in *Handbook of Socialization Theory and Research* (D. A. Goslin, ed.), Rand-McNally, Chicago, 1969.
33. J. Piaget, *The Child's Conception of the World*, Routledge & Kegan Paul, London, 1929.
34. J. Piaget, *The Moral Judgment of the Child*, Routledge & Kegan Paul, London, 1932.

35. E. Turiel, *The Development of Social Knowledge: Morality and Convention*, Cambridge University Press, Cambridge, England, 1983.
36. R. DeVries and A. Goncu, Interpersonal relations in four-year dyads from constructivist and Montesorri programs, *J. Appl. Dev. Psychol.* 8, 481–501 (1987).
37. B. Friedman, Societal issues and school practices: An ethnographic investigation of the social context of school computer use, paper presented at the annual meeting of the American Educational Research Association, Boston, April 1990 (ERIC Document Reproduction Service No. ED 321 740).
38. L. P. Nucci and M. Nucci, Children's responses to moral and social conventional transgressions in free-play settings, *Child Dev.* 53, 1337–1342 (1982).
39. R. DeVries, Children's conceptions of shadow phenomena, *Gen. Soc. Gen. Psychol. Monographs* 112, 479–530 (1986).
40. L. P. Nucci and E. Turiel, Social interactions and the development of social concepts in preschool children, *Child Dev.* 49, 400–407 (1978).
41. G. B. Saxe, *Culture and Cognitive Development: Studies in Mathematical Understanding*, Lawrence Erlbaum Press, Hillsdale, New Jersey, 1990.
42. L. A. Suchman, *Plans and Situated Actions: The Problem of Human–Machine Communication*, Cambridge University Press, Cambridge, England, 1987.
43. C. Allen and R. Pea, Reciprocal evolution of research, work practices and technology, in *Proceedings from the Conference on Participatory Design 1990* (A. Namioka and D. Schuler, eds.), Computer Professionals for Social Responsibility, Palo Alto, 1990.